Predicción del mantenimiento en entornos de cuarta generación

Antonio Martínez[a,b], Mario Piattini[a]

[a]*Universidad de Castilla-La Mancha*

[b]*Excma. Diputación de Ciudad Real*

Resumen

Los entornos de cuarta generación sustituyen cada vez más a los lenguajes de tercera generación como plataforma de desarrollo habitual de sistemas informáticos, por lo que se hace imprescindible controlar su complejidad y facilidad de mantenimiento. Un aspecto importante al desarrollar aplicaciones software es poder predecir y controlar el tiempo de mantenimiento de estas aplicaciones. Una forma de realizar este control es mediante la utilización de métricas específicas para estos entornos, campo que ha recibido poca atención dentro de la ingeniería del software.

En este trabajo se desarrolla, utilizando el método de análisis de regresión lineal, un modelo de predicción para estimar el tiempo de mantenimiento de una aplicación compuesta de 143 programas escritos en un entorno de cuarta generación, que utilizan preferentemente lenguaje SQL en modo consulta. El modelo desarrollado tiene una exactitud de MMRE=23,687% y PRED(0,25) = 0,825.

1. Introducción

Muchas organizaciones que utilizan sistemas de información basados en lenguajes de tercera generación, como COBOL, están evolucionando sus sistemas a entornos de cuarta generación, que son más efectivos y utilizan técnicas modernas facilitando el mantenimiento. Estas herramientas han recibido varios nombres, entre otros, entornos de cuarta generación, lenguajes de cuarta generación (4GL), generadores de aplicaciones, o sistemas de cuarta generación (4GS) [1]. Debido a su creciente difusión, se hace cada día más imprescindible contar con un conjunto de métricas que permita asegurar la calidad de los sistemas desarrollados con este tipo de entornos; especialmente la facilidad de mantenimiento, ya que el mantenimiento representa el mayor problema del desarrollo software suponiendo entre el 60 y el 80 por ciento de los costes del ciclo de vida [2], [3]. Las métricas del software son un buen medio para entender, monitorizar, controlar, predecir y probar el desarrollo software y los proyectos de mantenimiento [4] y también pueden ser utilizadas para que los profesionales e investigadores puedan tomar mejores decisiones [5].

En este trabajo se propone un modelo para la predicción del mantenimiento de programas desarrollados en lenguaje de cuarta generación que utilizan preferentemente el lenguaje SQL (Los comités de estandarización ISO/IEC y ANSI recientemente propusieron una nueva versión del SQL conocido como SQL:1999 (previamente SQL3) [6]. Este trabajo está enfocado en la versión (SQL'89)) en modo consulta (sentencia SELECT). En la sección 2 se describen las propuestas para medir programas escritos en lenguaje de cuarta generación. En la sección 3 se presentan los detalles del estudio empírico. En la sección 4 se muestra los resultados y el análisis de los datos. Finalmente, en la sección 5 se perfilan las conclusiones y trabajos futuros.

2. Métricas para el sublenguaje de manipulación de base de datos

Se han definido diferentes tipos de métricas para 4GL. Hasta el momento algunos trabajos se han llevado a cabo para estimar el esfuerzo de desarrollo y la correlación del mismo con el tamaño de un programa [7], [8], [9] describen un estudio empírico para predecir el tamaño de sistemas 4GL basado en varias métricas derivadas de los diagramas ER. La intención de nuestro trabajo es desarrollar un sistema de predicción útil para estimar el tiempo de mantenimiento de aplicaciones de software desarrolladas en entornos de cuarta generación. Para ello hemos identificado en los entornos 4GL distintos sub-lenguajes: de control procedimental, de control visual, de manejo de excepciones, de definición de base de datos, de manipulación de base de datos, de control de seguridad y de control de transacciones, [10]. A continuación, proponemos tres métricas para el sub-lenguaje de manipulación de base de datos y las particularizamos para la sentencia SELECT.

Métrica NT

Expresa el número de tablas que contiene la sentencia SELECT.

Métrica NA

Número de anidamientos de la sentencia SELECT, considerando la propia sentencia como un anidamiento SELECT. Por tanto si la sentencia SELECT no tiene anidados NA=1, si tiene un SELECT NA=2 y así sucesivamente.

Métrica A

Dentro de la sentencia SELECT señala si hay agrupamiento (A = 1) o no lo hay (A = 0).

3. Detalles del estudio

3.1. Características generales del sistema

El sistema está compuesto de una aplicación con 143 programas que gestiona el Departamento de Informática de la Excma. Diputación de Ciudad Real. Los programas se desarrollan utilizando prototipos, entrevistándose con el usuario alrededor de cuatro ocasiones durante un periodo de cinco semanas. La aplicación es un sistema de procesamiento de transacciones de consultas y realiza actividades de mantenimiento de los datos estando escrita en lenguaje de cuarta generación que utiliza sobre todo el lenguaje SQL en modo consulta, para satisfacer los requerimientos reales del servicio de personal de la propia entidad.

Los programas son todos de pequeño a mediano tamaño, como se ilustra en la siguiente escala de valores: cada programa incluye como puede verse en la tabla 1, una media de 6 tablas, 2 anidamientos y 1 agrupamiento.

Uno de los aspectos positivos de la aplicación es que fue construida completamente por el mismo grupo de desarrollo, empleándose la misma metodología y utilizando la misma herramienta, el CA-OpenIngres/4GL. Estos factores pueden ser considerados como una constante en el análisis y por lo tanto el grado de fiabilidad del estudio puede ser alto [11].

4. Resultados y analisis de los datos

4.1. Estadística descriptiva

La estadística general descriptiva para cada una de las variables se muestra en la tabla 1.

Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. típ.	Varianza	Asimetría	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico
NT	143	22	0	22	6,04	5,51	30,364	1,361	,203
NA	143	6	0	6	1,92	1,85	3,438	,712	,203
A	143	1	0	1	,40	,49	,241	,419	,203
TIEMPO	143	289	1	290	74,36	76,41	5838,965	1,205	,203
N válido (según lista)	143								

Tabla.1 Estadísticos descriptivos para cada medida

4.2. Análisis de Correlación

Para el test de correlación se utilizan los coeficientes estadísticos de Pearson y la correlación no paramétrica de Spearman tanto como para identificar potencialmente las relaciones entre la variable dependiente tiempo de mantenimiento(TIEMPO) y las variables

independientes (NT,NA,A) así como, las interrelaciones que puedan existir entre las variables. El resultado se muestra en las tablas 2 y 3.

Ambos conjuntos de correlación estadística evidencian fuertes relaciones significativas entre la especificación de la variable tiempo de mantenimiento y las métricas definidas. En particular, las relaciones entre la especificación de las medidas NT, NA y A y la variable TIEMPO de mantenimiento son significativas con la estadística más conservativa de Spearman.

Correlaciones

		NT	NA	A	TIEMPO
Correlación de Pearson	NT	1,000	,796**	,280**	,986**
	NA	,796**	1,000	,258**	,881**
	A	,280**	,258**	1,000	,284**
	TIEMPO	,986**	,881**	,284**	1,000
Sig. (bilateral)	NT	.	,000	,001	,000
	NA	,000	.	,002	,000
	A	,001	,002	.	,001
	TIEMPO	,000	,000	,001	.
N	NT	143	143	143	143
	NA	143	143	143	143
	A	143	143	143	143
	TIEMPO	143	143	143	143

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla.2 Coeficientes de correlación de Pearson

Correlaciones

			NT	NA	A	TIEMPO
Rho de Spearman	Coeficiente de correlación	NT	1,000	,799**	,286**	,964**
		NA	,799**	1,000	,243**	,908**
		A	,286**	,243**	1,000	,283**
		TIEMPO	,964**	,908**	,283**	1,000
	Sig. (bilateral)	NT	.	,000	,001	,000
		NA	,000	.	,003	,000
		A	,001	,003	.	,001
		TIEMPO	,000	,000	,001	.
	N	NT	143	143	143	143
		NA	143	143	143	143
		A	143	143	143	143
		TIEMPO	143	143	143	143

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla.3 Coeficientes de correlación de Spearman

4.3. Análisis de regresión

La regresión lineal (el software utilizado para el cálculo de la ecuación del modelo de regresión lineal es el SPSS/PC+ 7.5).

es una de las técnicas clásicas para analizar y construir modelos de estimación. Se asume que la variable dependiente está linealmente relacionada con las variables independientes.

El método de análisis de regresión utilizado es el de pasos sucesivos, que consiste en ir introduciendo en distintas etapas una variable independiente distinta. La primera variable que entra en el modelo es la más correlacionada con la dependiente, en este caso es NT, con coeficiente de correlación de Pearson de 0,986. Es la variable independiente que explicará un porcentaje máximo de la dependiente (TIEMPO).

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	NT		Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).
2	NA		Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).

a. Variable dependiente: TIEMPO

Tabla.4 Variables introducidas/eliminadas

Las siguientes variables que van a ir entrando en cada paso ya no van a depender del coeficiente de correlación con la dependiente, sino que van a depender de la correlación parcial y la tolerancia. En el primer paso han quedado fuera las variables restantes, NA, A como puede verse en la tabla 5. La columna de *beta dentro* como puede verse en la tabla 5 nos proporciona los coeficientes tipificados que tendrían estas variables en el modelo de regresión si fuesen incluidas en la etapa siguiente. La columna significación como puede verse en la tabla 5 (columna Sig) nos va a indicar la variable que va a entrar en el siguiente paso, para este experimento el sig tomado es de 0,05, por tanto la que tenga significación menor que 0,05. En la tabla 5 observamos que NA que tiene significación de 0,000 mientras que A tiene una significación de 0,561. Por tanto como puede verse en la tabla 4 la variable independiente NA entra en la segunda etapa.

Variables excluidas[c]

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad: Tolerancia
1	NA	,261[a]	34,075	,000	,945	,366
	A	,009[a]	,582	,561	,049	,922
2	A	-,001[b]	-,290	,772	-,025	,918

a. Variables predictoras en el modelo: (Constante), NT

b. Variables predictoras en el modelo: (Constante), NT, NA

c. Variable dependiente: TIEMPO

Tabla.5 Variables excluídas

Para saber si tiene sentido en este estudio la regresión necesitamos el análisis de varianza de regresión. En esta tabla 6 observamos un valor de F de 4909,152 y un valor de sig=0 para la primera etapa, lo que quiere decir que la regresión es significativa para cualquier nivel de significación. En la segunda etapa tenemos un valor de F de 23.213,01 con una significación de sig=0 por lo que igualmente la regresión es significativa para cualquier nivel de significación.

ANOVA [c]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	805983,74	1	805983,7	4909,152	,000[a]
	Residual	23149,355	141	164,180		
	Total	829133,09	142			
2	Regresión	826642,24	2	413321,1	23231,01	,000[b]
	Residual	2490,850	140	17,792		
	Total	829133,09	142			

a. Variables predictoras: (Constante), NT

b. Variables predictoras: (Constante), NT, NA

c. Variable dependiente: TIEMPO

Tabla.6 Análisis de Varianza

Existe una fuerte correlación entre la tres variables definidas en el modelo de regresión lineal múltiple, para determinar el tiempo de mantenimiento desde la especificación de las medidas, ilustradas por los valores de la estadística de correlación de Pearson como se muestra en la tabla 2.

El modelo de regresión lineal múltiple (método paso a paso) aplicado a la aplicación compuesta de 143 programas (observaciones) da el siguiente modelo de regresión como puede verse en la tabla 7:

TIEMPO= -11,513 + 10,789 (NT) + 10,758 (NA)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-8,243	1,593		-5,174	,000
	NT	13,672	,195	,986	70,065	,000
2	(Constante)	-11,513	,533		-21,593	,000
	NT	10,789	,106	,778	101,571	,000
	NA	10,758	,316	,261	34,075	,000

a. Variable dependiente: TIEMPO

Tabla.7 Coeficientes

Este modelo tiene un R^2 ajustado de 0,997 como se ve en la tabla 8, indicando que explica el 99% de la varianza en la implementacion de la variable tiempo de mantenimiento. El valor de R^2 da una medida de la consistencia de un modelo de regresión específico Este fue un resultado interesante, sugiriendo que se puede predecir en un grado significativo el tiempo de mantenimiento de la aplicación a partir de su finalización, considerado desde la puesta en marcha hasta el momento actual (Un año después de la puesta en marcha de la misma).

Resumen del modelo[c]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,986[a]	,972	,972	12,81	
2	,998[b]	,997	,997	4,22	2,037

a. Variables predictoras: (Constante), NT

b. Variables predictoras: (Constante), NT, NA

c. Variable dependiente: TIEMPO

Tabla.8 Resumen del modelo

Una cuestión que puede aumentar la aceptabilidad del modelo, es la inclusión de dos términos "independientes" que en realidad están interelacionados (como se ilustra en la tabla de correlaciones presentada anteriormente). Los modelos qué presentan términos relacionados pueden ser inestables o menos fácil de generalizar a otros conjuntos de datos [12], [13]. Se puede crear otro modelo de regresión lineal que sólo incluye la variable independiente más significativa. La forma de este modelo como puede verse en la tabla 7 se muestra a continuación:

TIEMPO= -8,243 + 13,672 (NT)

Este modelo tiene asociado un R^2 ajustado de 0,972 como se puede ver en la tabla 8, un valor menor comparado con el modelo de dos variables.

Puede afirmarse que deberían incluirse ambos términos (NT y NA) pero existe una fuerte correlación entre ambas (0,796 Pearson y 0,799 Spearman). Una técnica que puede ayudar a determinar el aumento asociado con la inclusión de términos extras en un modelo de regresión es el test de R^2 adecuado [14]. Un subconjunto de variables predictoras debe cumplir:

$$R^2_{sub} > 1 - (1- R^2_{full}) (1 + d_{n,k})$$

donde:

R^2_{sub} es el valor alcanzado de R^2 con el subconjunto de variables predictoras.

R^2_{full} es el valor alcanzado de R^2 con todas las variables predictoras.

$d_{n,k} = (k*F_{k,n-k-1})/n-k-1$

donde:

k = el número de variables predictoras en el modelo

n = el número de observaciones de la muestra

F = la F statistic para α para n,k grados de freedom

En esto caso:

k=2, n=143, α=0,05

R^2_{full} = 0,997 (para el modelo con dos variables predictoras)

R^2_{sub} = 0,972 (para el modelo con una variable predictora)

$\Rightarrow d_{143,2} = (2*F_{2,140})/143-2-1 = (2*2,99)/140$

$d_{143,2} = 0,03126$

$\Rightarrow R^2_{sub} > 1 - (1- R^2_{full}) (1 + d_{n,k}) = 1-(1-0,997) (1+0,042) = 1-(0,03) *(1,042)$

$= 0,968$

Dado que el valor de R^2 del modelo de una única variable predictora 0,972 es mayor que el umbral mínimo del valor de R^2 adecuado para el modelo completo (dos variables predictoras) (0,968), podemos decir que el modelo incluyendo solo el término NT es tan efectivo en términos de consistencia como el modelo conteniendo ambas variables NT y NA. Además adoptando el modelo de una sola variable predictora nos puede ayudar a superar algunos problemas referidos anteriormente en relación a incluir términos predictivos relacionados en un modelo de regresión.

Un aspecto importante y deseable del modelo de regresión es la consistencia que viene ilustrada por R^2 pero no es suficiente únicamente esta característica. En realidad, un modelo

puede ser muy consistente, pero excesivamente inseguro, en términos de la correspondencia de los valores de los pares de la muestra. Dos indicadores utilizados comúnmente para evaluar la capacidad de predicción de un modelo son la magnitud media del error relativo (MMRE) y el indicador pred (Nivel de Predicción l (PRED(l), donde l es un porcentage, se define como el cociente del número de casos en los cuales las estimaciones estan dentro del límite absoluto l de los valores actuales dividido por el número total de casos) [15].

La magnitud de error relativa (MRE) es una medida normalizada de la discrepancia entre los valores actuales (V_A) y los valores estimados (V_F):

$$MRE = Abs(V_A - V_F)/ V_A$$

La media MRE (MMRE) es la media de los valores para este indicador sobre todas las observaciones de la muestra. Un valor bajo para MMRE generalmente indica un modelo más exacto.

La media PRED da una indicación de todos los valores estimados para un conjunto de datos, basados en el valor MRE para cada par de datos:

$$Pred(l) = i/n$$

donde:

l es el valor umbral seleccionado para MRE

i es el número de par de datos con MRE menor o igual que l

n el el número total de pares de datos.

En nuestro ejemplo, estas dos magnitudes tienen los siguientes valores:

MMRE=23,687% y pred(0,25) = 118/143=0,825, significando que el 82,5% de los valores estimados caen dentro del 25% de sus correspondientes valores actuales.

4.4. Análisis del Residual

El valor del coeficiente Durbin-Watson tiene un valor de 2,037 deduciéndose la incorrelación de los residuos, y por tanto, de las observaciones como se puede ver en la tabla 8.

4.5. Observaciones extremas significativas (Outliers)

Encontramos en el conjunto de datos un pequeño número de puntos extremos. En particular, hay observaciones que tienen valores extremos significativos (outliers) con un MRE asociado de 3,30 bajo el modelo de las dos variables. Este valor se encontró en los

asociado un tiempo de mantenimiento.

5. Conclusiones y trabajo futuros

Cómo es lógico, no podemos generalizar este modelo a otros entornos de cuarta generación ni si quiera al mismo entorno ya que pueden existir instalaciones que no tengan el mismo tipo de aplicaciones (en las que predomine el lenguaje SQL de consultas). A pesar de eso creemos interesante investigar sobre modelos que puedan predecir el esfuerzo de mantenimiento es estos nuevos entornos de cuarta generación.

Hay una gran necesidad de medir la calidad de las aplicaciones basadas en lenguajes de cuarta generación. Las medidas pueden ayudarnos para capturar ciertos atributos de calidad de software y estos atributos se pueden utilizar para construir mejores productos software [16].

Este trabajo muestra que es posible estimar el tiempo de mantenimiento de aplicaciones desarrolladas en entornos de cuarta generación, que utilizan el lenguaje SQL en modo consulta (sentencia SELECT), utilizando simplemente una herramienta que cuente el número de tablas y el número de anidamientos de cada programa que componen la aplicación.

En particular, encontramos que el número de tablas y el número de anidamientos son entradas para el sistema de predicción derivado del análisis de regresión de nuestro conjunto de datos.

Agradecimientos

Este trabajo forma parte del proyecto MANTICA, parcialmente financiado por la CICYT y la Unión Europea (1FD97-0168).

Referencias

[1] Holloway, S. (ed.). *Fourth-Generation Systems, their scope application and methods of evaluation.* Chapman and Hall,1990.

[2] Card, D.N. y Glass, R.L. *Measuring Software Design Quality.* Englewood Cliffs, 1990.

[3] Pigoski, T.M. *Practical Software Maintenance.* Wiley Computer Publishing, 1997.

[4] Briand, L.C., Morasca, S. y Basili, V., "Property-based software engineering measurement", *IEEE Transactions on Software Engineering,* vol. 22, nº1, 1996, pp.68-85.

[5] Pfleeger.S.L., "Assessing Software Measurement", *IEEE Software*. Marzo/Abril, 1997, pp.25-26.

[6] Eisenberg, A. y Melton, J., "SQL:1999, hasta ahora conocido como SQL3", *SIGMOD Record*, vol.28, nº1, 1999, pp.131-138.

[7] Dolado, J.J., "A Study of the Relationships among Albrecht and Mark II Function Points, Lines of Code 4GL and Effort", *J. Systems Software*, vol.37, 1997, pp.161-173.

[8] Verner J. y Tate G., "Estimating Size and Effort in Fourth-Generation Development", *IEEE Trans. On Software Engineering*, 1988, 15-22.

[9] Bourque, P. y Côté, V., "An experiment in software sizing with structured analysis metrics", *Journal of Systems and Software* vol. 15, 1991, pp.159-172.

[10] Martínez, A. y Piattini, M., "Validation of 4GL metrics", *Proc.of the Software Measurement in Practice, 10th Anniversary Conference*. United Kingdom Software Metrics Association, octubre, 1998, pp. 1-19.

[11] MacDonell, G., Shepperd y J., Sallis, J., "Metrics for Database Systems: An Empirical Study", *IEEE Software*, 1997, pp. 99-107

[12] Coupal, D. y Robillard, P.N., "Factor analysis of source code metrics", *Journal of Systems and Software*, vol.12 , 1990, pp.263-269.

[13] Kitchenham, B.A. y Pickard, L.M. "Towards a constructive quality model. Part II: Statistical techniqyes for modelling software in the ESPRIT REQUEST project", *Software Engineering Journal* , julio, 1987, pp.114-126

[14] Neter,J., Wasserman, W. y Kutner, M.H. *Applied Linear Regression Models*. Irwin: Homewood IL, 1983.

[15] Dolado, J.J. , Fernandez, L., Otero, MC. y Urkola, L., "Software effort estimation: the elusive goal in project management",. *ICEIS*,1999, pp.412-418.

[16] Zuse, H. *A Framework of Software Measurement* , Ed. Walter De Gruyter.1998